**Developing a Recommendation System for Retailers**

Anika Jannat Rimu   
The University of Texas at Arlington   
May 2022

**Abstract**

This project develops a recommendation system containing 12 articles for 7 days for H&M. A collaborative- based recommendation system was developed using the customer, article and training dataset. Exploratory Data Analysis was done to understand the data. Once the key features were identified, the data was cleared and standardized. The recommendation system was developed using K-Nearest Neighbor and Neural Network algorithm. While the KNN model is a classification model, Neural Network is a regression model so Root Mean Squared error was used to improve the accuracy of the regression model. The recommendations were different from each other so future studies can be done on the customer feedback, sales and profit after implementing the recommendation system as there is no other way to determine the accuracy.

**1 Introduction**

This problem is a recommendation system problem that is organized by the clothing brand H&M. A recommendation system is a subclass of machine learning that can predict customer preference by using details of the individual user or the community of other users who have similar likings as the individual user (Crespo et al., 2011). Recommendation systems can be developed using two types of algorithms- content based recommendation system and collaborative recommendation system (Adomavicius & Tuzhilin, 2005; Balabanovic & Shoham, 1997). Content based algorithm relies on information about the product and recommends product similar to the one he has liked in the pard. This type of recommendation system is widely used in recommending documents, web pages, publications, jokes or news. Collaborative recommendation system on the other hand takes account of the users and the similarity in their likings. This kind of system uses users’ historical preferences and ratings to recommend a set of products. Big companies such as Amazon, Facebook, Twitter even dating apps uses this type of system to generate recommendations (Fatourechi, 2015).

Collaborative recommendation system has recently gained a lot of popularity among the online retailers (Mild & Reutterer, 2003). Several research has shown sales to increase after implementing a recommendation system. Pathak et al., (2010) has found recommendations not only to increase sales but also to add flexibility to the retailers to adjust their price. By implementing digital word of mouth and recommendations, retailers can charge higher price while increasing the demand at the same time by giving their customers access to more information regarding quality and match. A case study done by Maddodi & Prasad (2019) shows that after Netflix started implementing a recommendation system to recommend their users older movies they might like, the demand for the newer superhit movies went down by 20% at that time. Amazon similarly started their item based collaborative filter back in 1998 and they have ever since surpassed their competitors like Alibaba and has become the biggest e-commerce company in the world (Smith & Linden, 2017). Amazon’s extremely powerful recommendation system contributes to 35% of its revenue. Other than these two companies some big online platforms such as Facebook, YouTube, TikTok etc. use powerful algorithm to get more users and. In this era, when all the information is in the palm of hand, how quickly that information can be provided to its target user can make all the differences when it comes to profit and sales.

Shafer et al. (1999) mentions that there are three ways recommendation systems enhance the sales- it can help customers find the product they wish to buy; it can increase the order size by recommending products throughout the shopping process especially during the checkout prices and lastly, it can gain customer loyalty by creating a value-added relationship between the site and the customer. Therefore, big companies such as H&M are investing on learning more about their customers so they can stay ahead of their competitors. Similar to the competition hosted by Netflix back in 2006 to improve their recommendation system- “Cineplex”, H&M also hosted a competition to develop a product recommendation system based on their data from previous transactions, customer and product meta data. By recommending clients the accurate product, H&M not only wants to increase their sales, btu also enhance the customer shopping experience by helping them make the right choice. This can in turn, also reduce less customer return and as a result, reduce emissions from transportations.

The goal of this project is to develop a collaborative recommendation system using both the previous transaction data and the product and customer meta data. Recommendation for 12 articles were made first using k-Nearest Neighbors method and then using Neural Network method. It is hypothesized that both recommendation systems will be able to predict accurate articles for the specific user.

**2 Data Exploration**

There were three datasets provided by H&M for this study. There was a customer metadata that included the date, customer id, article id, price, number of purchases, their club member status, fashion frequency, age and postal code. The article metadata consisted of several information about the article including article name, graphical appearance etc. Lastly, there was training dataset that consisted of date of purchase, customer id, article id and price of the product.

The first thing that was done for data processing was merging the training data with the customer data by using Panda’s “merge” command. This allowed the customer id and the respective article id, price and number of purchases to be added to the customer metadata. This way the training data was associated with the customer data to get more knowledge on customer behavior. The data was then cleaned by deleting rows that didn’t have any value for age, Active status and Fashion News status. Then any cell that had no value was replaced with 0. The age of the customers was placed in bins containing 10 numbers for example 0-9, 10- 19 etc.

To identify the features for the classification model, a detailed Exploratory Data Analysis (EDA) was performed. The first set of EDA was performed on the sale of products (Figure 1). Next the relationship between number of customers, number of articles purchased by each customer, Fashion News Frequency, Club Member status, age and sales channel ID (1 for in store 2 for online) were analyzed (Figure 2-3).

From this EDA, the age of the customers, their fashion frequency, club member status, and price of the product were used as features. One hot encoding was done on the age bins, club member status and fashion news frequency to convert them to categorical variables. The features were then standardized and transformed using MinMaxScaler of the sklearn.preprocessing package.

Chart, bar chart

Description automatically generatedChart, bar chart

Description automatically generated

Sales Quantity

Sales Quantity

Figure 1: Number of sales by index group name (left) and garment group name (right)

Chart, pie chart

Description automatically generatedChart, pie chart

Description automatically generated

Figure 2: Pie chart of fashion frequency (left) and club member status of customers (right)

Chart, bar chart

Description automatically generated Chart, bar chart

Description automatically generated

Figure 3: Number of articles purchased per age group and the sales channel preferred

**3 Prediction Methods**

1. K- Nearest Neighbor Algorithm:

In order to find the recommendation, an unsupervised model was used since there is no specific label that the model can identify, instead it needs to analyze the pattern to find the hidden features among various users. The NearestNeighbors function from Scikit-learn was used with n\_neighbors as 12 for 12 recommendations, radius= 1, metric= minkowski and p=2.

Using this algorithm, the distances and the id list of the distances were obtained. This determined the nearest neighbor of each row. Next a function was made for the recommendation system. In this function the index of the specific customer is first identified. Once the index is identified, the corresponding id list generated from the nearest neighbor algorithm is identified. Next, based on that id, the corresponding 12 neighbors of article list is created and recommended to the user. Lastly the function is called using a customer ID and the list is printed out.

The accuracy of the KNN method can only be determined by taking customer feedback as there is no other way to predict what the customer might have thought of purchasing.

1. Neural Network Algorithm:

A neural network algorithm was developed to make interaction between two input variables to create a matrix factorization table. Since age has been shown as an important factor in article selection, it was used as the output variable that is determined from the interaction of the two input variables – customer id and article id. This can be considered a regression problem since age is being targeted and getting predicted by it using customer and article interaction.

First the unique number of customers and articles are determined and indexed starting at 0 using a dictionary. These indexes are added as separate columns to dataset for mapping. There were a total of 2950 unique users, 4905 unique articles and the minimum age was 18 and the maximum age was 78. Age was then standardized between 0 to 1. The model is then split into training and testing set with 90% being training set.

Each variable i.e., customer and article are then embedded to 50 dimensional latent feature space using Keras embedding function. It is then flattened so it can create a 50-dimensional vector. A dot product of both flattened vectors of customers and articles is then calculated. The dot product is then passed to the multiple dense layers. Dense layers are the multilayer perceptron (MLP) i.e, a class of feedforward artificial neural network. Dense one here consists of 150 neurons, dense two consists of 50 neurons and dense 3 is output dense layer that is scalar and consists of one single neuron since one value needs to be predicted i.e., age (Figure 4). The model is then created and compiled consisting of the two inputs and the predicted output. The optimizer ‘Adam’ is used and mean squared error is also calculated since it is a regression problem. Next, the training is done using epoch= 40 i.e., number of cycles of backward and forward passes. 40 was used because the curve for loss became stable after 40 (Figure 5). The rest of the model separates each user’s articles bought from articles not bought and makes predictions on items that is not previously purchased by the user.

Diagram

Description automatically generated

Diagram

Description automatically generated

Figure 4: Collaborative Neural Network (Kumar, 2021)

**Chart, histogram

Description automatically generated**

Figure 5: Root mean squared error

**4 Results**

Recommendation from KNN model:

Text

Description automatically generated

Recommendation from Neural Network model:

Text, letter

Description automatically generated

There are differences seen in both methods. There is no way to know which method is correct as both has different advantages and disadvantages. KNN here is an unsupervised classification method that is relying on distances, but it takes account of similarities in several factors; whereas the Neural Network model here is a regression model that has been extensively trained but it only accounts for the age as an interaction factor between the customers and the articles. This model also takes account of root mean square error values which are very low, so the accuracy is higher but, in the end, it only considers one factor. The only way the accuracy of either model can be analyzed is by using both models and analyzing the data for how each model affect the sales.

**5 Conclusion**

This project builds a recommendation system to analyze H&M dataset and recommend 12 articles for each customer. It uses two methods to build the system- K-Nearest Neighbor and Neural Network. For KNN, the Minkowski distance is used and features such as age, fashion news frequency, club member status and price are used to calculate the distance. For Neural Network, matrix factorization is done by finding the interaction between customers and articles using age. The recommendations obtained from both methods were different than each other but the accuracy of each method cannot be determined. This study can be improved by trying out other methods such CNN on the article images for a content-based recommendation system or using a hybrid system. Future studies could include a case study of changes in sales and profit after implementing each recommendation system.

**6 Acknowledgement**

This project is to fulfill a requirement of IE 6318 Data Mining and Analytics course in the University of Texas at Arlington. I would like to thank Dr. Wang and TA Linh for their continuous support.

**7 References**

Adomavicius, G., & Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE transactions on knowledge and data engineering*, 17(6), 734-749.

Crespo, R. G., Martínez, O. S., Lovelle, J. M. C., García-Bustelo, B. C. P., Gayo, J. E. L., & De Pablos, P. O. (2011). Recommendation system based on user interaction data applied to intelligent electronic books. *Computers in human behavior*, *27*(4), 1445-1449.

Fatourechi, M. (2005). The Evolving Landscape Of Recommendation Systems. Retireved from <https://techcrunch.com/2015/09/28/the-evolving-landscape-of-recommendation-systems/#:~:text=Collaborative%20Filtering&text=Companies%20that%20employ%20this%20model,Google%20News%20and%20Last.fm>. Accessed on 05/07/2022

Kumar, A. (2021). Recommendation System Neural Network. Retrieved from <https://www.youtube.com/watch?v=x6sTjmPIXr0&t=684s>. Accessed on 05/09/2022

Maddodi, S. (2019). NETFLIX bigdata analytics-the emergence of data driven recommendation. *Srivatsa Maddodi, & Krishna Prasad, K.(2019). Netflix Bigdata Analytics-The Emergence of Data Driven Recommendation. International Journal of Case Studies in Business, IT, and Education (IJCSBE)*, *3*(2), 41-51.

Mild, A., and Reutterer, T. (2003). An improved collaborative filtering approach for predicting cross-category purchases based on binary market basket data. Journal of Retailing and Consumer Services, 10, 3 , 123–133.

Pathak, B., Garfinkel, R., Gopal, R. D., Venkatesan, R., & Yin, F. (2010). Empirical analysis of the impact of recommender systems on sales. *Journal of Management Information Systems*, *27*(2), 159-188.

Schafer, J. B., Konstan, J., & Riedl, J. (1999, November). Recommender systems in e-commerce. In *Proceedings of the 1st ACM conference on Electronic commerce* (pp. 158-166).

Shoham, Y. (1997). Combining content-based and collaborative recommendation. *Communications of the ACM*.

Smith, B., & Linden, G. (2017). Two decades of recommender systems at Amazon. com. *Ieee internet computing*, *21*(3), 12-18.